

The lifecycle of provenance metadata and its associated challenges and opportunities

Paolo Missier¹

School of Computing Science
Newcastle University
Newcastle upon Tyne, UK Paolo.Missier@ncl.ac.uk

Abstract. This chapter outlines some of the challenges and opportunities associated with adopting provenance principles [CFLV12] and standards [MGC⁺15] in a variety of disciplines, including data publication and reuse, and information sciences.

Keywords: Provenance data modelling, provenance lifecycle, provenance analytics

Using provenance in a broad diversity of application areas and disciplines entails a number of challenges, including specialising the generic provenance and domain-agnostic data model, PROV. This chapter provides a brief overview of these challenges, using the *provenance lifecycle* framework shown in Fig. 1 as a reference.

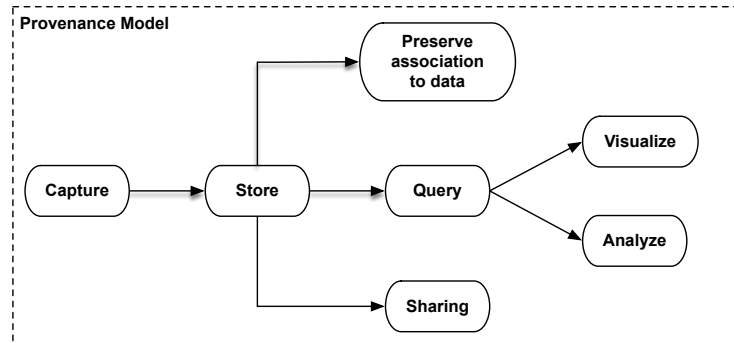


Fig. 1. Schematic of provenance lifecycle

1 Provenance definitions and model

PROV, the Provenance standard, is a family of specifications released in 2013 by the Provenance Working Group, as a contribution to the Semantic Web suite

of technologies at the World Wide Web Consortium. PROV aims to define a *generic* data model for provenance that can be extended, in a principled way, to suit many application areas. The PROV-DM document [MMB⁺12] provides an operational definition of provenance for the community to use and build upon:

Provenance is defined as a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing.

The document goes on to position the definition in the context of Information Management:

The provenance of information is crucial in deciding whether information is to be trusted, how it should be integrated with other diverse information sources, and how to give credit to its originators when reusing it. In an open and inclusive environment such as the Web, where users find information that is often contradictory or questionable, provenance can help those users to make trust judgements.

1.1 PROV as a community data model and ontology

The specifications define a data model and an OWL ontology, along with a number of serializations for representing aspects of provenance. The term *provenance*, as understood in these specifications, refers to information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness (PROV-Overview [w3c12]). The specifications include a combination of W3C *Recommendation* and *Note* documents. Recommendation documents include (i) the main PROV data model specification (PROV-DM [MMB⁺12]), with an associated set of constraints and inference rules (PROV-CONSTRAINTS [CMM12]); (ii) an OWL ontology that allows a mapping of the data model to RDF (PROV-O [LSM⁺12]), and (iii) a notation for PROV with a relational-like syntax, aimed at human consumption (PROV-N [MMCSR12]). All other documents are Notes. These include PROV-XML, which defines a XSD schema for XML serialization [pro13c]. PROV-AQ, the Provenance Access and Query document [MHS⁺12], which defines a Web-compliant mechanism to associate a dataset to its provenance; PROV-DICTIONARY [pro13b], for expressing the provenance of data collections defined as sets of key-entity pairs; and PROV-DC [pro13a], which provides a mapping between PROV-O and Dublin Core Terms.

1.2 The provenance of PROV

PROV is the result of a long incubation process within the provenance community. The idea of a community-grown data model for describing the provenance of data originated around 2006, when consensus began to emerge on the benefits of having a uniform representation for data provenance, process documentation,

data derivation, and data annotation, as stated in [MCF⁺11]. The First Provenance Challenge [MLAB08] was then launched, to test the hypothesis that heterogeneous systems (mostly in the e-science / cyberinfrastructure space), each individually capable of producing provenance data by observing the execution of data-intensive processes, could successfully exchange such provenance observations with each other, without loss of information. The Open Provenance Model (OPM) [MCF⁺11] was proposed as a common data model for the experiment. Other Provenance Challenges followed, to further test the ability of the OPM to support interoperable provenance.

In September, 2009, the W3C Provenance Incubator Group was created. Its mission, as stated in the charter [pro09], was to “provide a state-of-the art understanding and develop a roadmap in the area of provenance for Semantic Web technologies, development, and possible standardization.” W3C Incubator groups produce recommendations on whether a standardization effort is worth undertaking. Led by Yolanda Gil at University of Southern California, the group produced its final report in Dec. 2010 [pro10]. The report highlighted the importance of provenance for multiple application domains, outlined typical scenarios that would benefit from a rich provenance description, and summarized the state of the art from the literature, as well as in the Web technology available to support tools that exploit a future standard provenance model. As a result, the W3C Provenance Working Group was created in 2011, chaired by Luc Moreau (University of Southampton) and Paul Groth (Vrije Universiteit Amsterdam). The group released its final recommendations for PROV in June, 2013.

1.3 Other notions of data provenance

Other formal models of data provenance exist, specifically in the context of database management. The provenance of a data item that is returned by a database query, for example, is defined by the semantics of the query itself, and mentions the fragments of the database state that were involved in the query processing [CCT09]. An algebraic theory in support of data provenance representation and management has been developed [Gre07]. This form of *fine-grained* provenance is often contrasted with *coarse-grained* provenance, which records the input / output derivations that are observed when functions are invoked, typically from within workflows and in the context of scientific data processing [DCBE⁺07]. Attempts have also been made to reconcile these two views, e.g., when declarative-style queries are embedded within procedural workflow processing [ADD⁺11].

2 Embracing provenance: status and opportunities

As illustrated in Fig. 1, there are a few key phases in the lifecycle of a provenance document: Production (Capture), persistent storage, Query, Sharing, Association with the underlying data products, and consumption/exploitation (Visualization/Analysis). The remainder of this short overview will only cover issues

concerning Capture, Storage and Query, and Analysis, using the following simple example to illustrate key issues in each of these phases.

In PROV, a provenance document is a set of assertions about the derivations that account for the production of a dataset, including, when available, its attribution. For example, one can use PROV to formally express the following facts:

Alice took draft v0.1 of paper P , made some edits during a certain time interval, and produced a new draft v0.2 of P .

In doing so, she used papers p_1 , p_2 as reference.

Alice then delegated Bob to do proofreading of P v0.2, which resulted in a new version v0.3 of P .

Alice also published a dataset D as supplementary material to P , which she has uploaded to a public data repository, for others to discover and reuse.

These facts can be expressed formally, using either RDF, XML, or PROV-N, the bespoke near-relational syntax mentioned earlier.

2.1 Extending PROV

The PROV Working Group worked hard to ensure that PROV can be extended in a principled way, in order to fit the needs of multiple disciplines where expressing the provenance of data may be important. Specifically, one can (i) use PROV-O, the PROV OWL ontology, in conjunction with other ontologies, in order to provide rich semantic annotations of data, and (ii) extend PROV-O itself with domain-specific provenance concepts.

As an illustration of (i), in the example above one can semantically characterize data products as “papers” of a certain type, along with the associated activities (editing, proofreading) using a suitable vocabulary, while at the same time characterizing their provenance using an RDF serialisation of the example statements above. As a reference, in the recent past we have demonstrated this capability in our specification of the *Janus* ontology [MSZ⁺10]. In brief, provenance and semantic annotations serve complimentary roles: the former tells the *history* of a data product, while the latter elucidates its *meaning*.

Regarding extending PROV, one notable example is the ProvONE ontology (formerly known as D-PROV) [MDB⁺13], aimed at capturing at the same time the data dependencies that emerge from observations during data creation (known as *retrospective* provenance), as well as the static structure of the process that is responsible for the generation of the process (known as *prospective* provenance) [LLCF10]. The latter is deliberately missing from PROV, owing to its generality. The D-PROV ontology specifically extends PROV to account for the structure of scientific workflows, a specific type of data-generating process that is important in many e-science applications.

In particular, the latest embodiment of D-PROV, called ProvONE [pro14], is currently in production use by the DataONE project (dataone.org). DataONE, a large NSF-funded project (2010-2018), is the largest Research Data conservancy project in the USA, with a focus on Earth Observational Data and ecology/climate data in particular. With a growing federation that already counts

tens of member repositories and hundreds of thousands of science data objects, the DataONE architecture places metadata indexing and management at the cornerstone of its data search and discovery capabilities. “Searching by provenance” is a new and unique feature that leverages the ProvONE data model, as well as the automated capture of retrospective provenance whenever R or Matlab (and, soon, Python) scripts that access DataONE science objects are executed.

The ProvONE ontology provides a template for extending PROV, which can be used in a number of other domains, as it illustrates proper use of the PROV extensibility points.

2.2 Provenance capture

Provenance is the result of observing a data transformation process in execution, including details of its inputs and outputs, be it a database query or a workflow, including processes carried out by humans or only partially automated. Key questions concerning the recording (“capturing”) of provenance include (i) what provenance-related events can be observed, (ii) what is the level of detail of these observations, and (iii) how does one deal with multiple, overlapping but inconsistent observations?

Regarding scientific data processing, the ability to record provenance relies entirely on the infrastructure on which the processes are executed. An increasing number of tools and systems are being retrofitted with provenance recording capabilities, including the best known workflow management systems [DCBE⁺07,MWHW13], and more recently, the Python [MBC⁺14] and the R languages [LB14b, LB14a] for data analytics. Two specific instances of provenance capture sub-systems for scientific workflows, that we have actively contributed to, are [MPB10], for the Taverna workflow management system developed in Manchester to support bioinformatics researchers [HWS⁺06,MSRO⁺10], and for the eScience Central workflow manager [HWWL11].

The case of completely automated processes which run in a centralized environment is, however, the simplest possible scenario. “Human-in-the-loop” processes are obviously more problematic, and are limited to capturing human interactions with information systems through a user interface. Clearly, solutions in this space are necessarily bespoke, with no known publications reporting specific case studies.

In each of these cases, the observations may be available at a specific level of abstraction, which may or may not be appropriate for the type of downstream analysis requirements (see below). These range from fine-grained, high-volume, system-level provenance (ie every file I/O operation in the system) [MCS11], to “coarse-grained” provenance from workflow executions, where only the inputs and outputs of each workflow block can be observed.

As a consequence of these varying levels of details, it becomes necessary to be able to adjust the quantity of information contained in a provenance document, i.e., by creating *views over provenance* that represent abstractions over provenance. In the example above, we could for instance conflate the editing and

proofreading activities into one, high-level “paper preparation” activity, and ignore the interim v0.2 of *P*. Our own work on *provenance abstraction* [MBG⁺14] builds upon prior research [BCBD07,DZL11], reflecting the user need not only to simplify the amount of provenance presented to the user, but also to *obfuscate* provenance in order to preserve its confidentiality.

A further complication in provenance capture, is that the observable processes normally take place on multiple, heterogeneous, autonomous and distributed systems, where the corresponding data is scattered. The provenance of an end data product must therefore be *reconstructed* by composing multiple, possibly inconsistent, and incomplete provenance fragments harvested from each of those systems. This is a relevant but under-studied area of research for provenance, with many potential applications that extend well beyond the realm of e-science.

2.3 Storage, Retrieval, and Query

Storing, indexing, and querying provenance documents requires a data layer not unlike that used to store the underlying data products that the provenance refers to. Data provenance that describes the history of large volumes of data is itself bound to have a high volume. Furthermore, if one includes in the provenance the intermediate data products that are generated as part of a complex data processing pipeline, it is easy to see that the size of the provenance documents may vastly exceed that of the data whose history it describes. Older and recent research has been devoted to studying the trade-offs between storing intermediate data products as part of provenance, which may incur a high storage cost [WHW15], as opposed to partially re-computing the data products (“smart rerun” [CBL11]).

Issues of dealing with large-scale provenance were addressed in the *BigProv* international workshop organized in 2013 and co-located with the EDBT conference. A number of submissions contributed to corroborate the hypothesis that the scalability of provenance management systems is becoming a practical problem if interesting analytics are to be derived from it. Amongst these, a study on reconstructing provenance from log files [GP13].

Provenance documents such as the one in our example are naturally expressed in the form of a graph. This suggests that graph databases (GDBMS) are suitable for their persistent storage, indexing, and querying. In our past work we have been experimenting with Neo4J, a new generation GDBMS, in order to study the scalability properties of provenance storage. In particular, we have developed ProvGen [FM14], a generator of synthetic provenance graphs of arbitrary size and with topology constraints. ProvGen is designed to create benchmarks for testing the performance of graph-based provenance data layers. It can generate provenance documents with millions of nodes and stores them in a Neo4J database.

At the same time, the standard RDF serialization of PROV, which specifies how provenance documents can be expressed using RDF triples that comply with the PROV ontology (PROV-O), lends itself well to storing provenance graphs

in existing RDF triple stores. However, despite the need for testing provenance data layers at scale, and our own past attempts at soliciting contributions that document scalability of provenance storage and query systems (the ProvBench workshop, co-located with *BigProv* (see above), to the best of our knowledge no official benchmarks have ever been released.

2.4 Provenance Analytics and novel uses for provenance

With the broad term “provenance analytics” we indicate all forms of consumption and exploitation of provenance corpora, once they have been captured and made available through suitable data engineering solutions, alluded to above. Relevant questions include: what can we learn from a large body of provenance metadata? what techniques and algorithms can be successfully borrowed from the realm of (Big) Data Analytics, in order to gain insight into data through its provenance?

Much has been made of provenance as a key form of metadata to help understanding the quality of data as well as its trustworthiness. A whole special issue of the ACM Journal of Data and Information Quality [Mis15a], has been devoted to the topic [Mis15b]. Despite several high quality submissions, however, more research is needed to fully elucidate the connection between data provenance and quality.

Many other opportunities are worth exploring that exploit provenance corpora in several domains. One line of research still in its infancy, concerns using provenance to ascribe *transitive credit* [Kat14] to scientists and other contributors who publish their datasets in public data repositories, for others to reuse. Data publication is a rapidly growing area of Open Science, which is based upon the assumption that scientists will spontaneously make their datasets public, as long as due credit is given to them through community mechanisms. Unfortunately, these mechanisms are still quite primitive, limited as they are to counting the number of citations to datasets, as they are found in paper publications (see for instance the *Making Data Count* project [KS15]). Instead, transitive credit pushes this embryonic notion of “credit for data” much further, as it leverages provenance to take into account multiple generations of data derivation and reuse.

Other disciplines farther away from computing and science will benefit from properly collected provenance, wherever providing accountability of a process execution is important. One example amongst many concerns *food safety*, where traceability of lots of food along a supply chain is critical to ensuring compliance with quality standards and proper handling, and to answer questions in case of accidents involving consumption of unsafe food.

2.5 Three key challenges for practical usability of provenance data

To conclude this overview, three areas when more research is needed in order to make provenance usable in practice are worth mentioning.

Incomplete and uncertain provenance. Generation and usage of data naturally occurs in many different ways through multiple, autonomous information systems. As a consequence, the provenance of such data is also naturally *fragmented and incomplete*. One major problem in provenance research is how to reconstruct a complete “big picture” out of such fragments. We are currently addressing this foundational problem in the specific setting of Open Research Data reuse, as this is a key issue when establishing transitive credit as mentioned above.

Trusted provenance. A second issue concerns accountability of the provenance documents themselves. To the extent that provenance documents are considered as a form of evidence for the underlying data, it is necessary to ensure that the provenance itself can be trusted not to have been tampered with. Using provenance traces in, say, a court of law, requires strong non-repudiability and integrity guarantees, which can only be provided by a trusted computing infrastructure [MMM05,LM10]. The notion of tamper-proof (or rather, tamper-evident) provenance has been touched upon in the past [ZCL09], but more research is needed as this clearly conflicts with the notion of provenance abstraction through views, alluded to above, namely when generating views involves *redacting* the provenance document itself [CKKT11].

Provenance to help the reproducibility of scientific processes. Lastly, we mention a long-standing promise on which provenance studies have largely yet to deliver. Much has been said (and there is no scope for a full survey here) of the role of provenance to support reproducible science, since the connection between reproducibility and provenance was first made back in 2008 [DF08].

Reproducibility is a known problem for a large number of scientific processes of the past, which are often encoded as a loose collection of scripts with external dependencies on ever-changing libraries, services, and databases. Practical solutions where provenance is used to ensure that these processes are reproducible are not readily available, however. In the recent past, we have addressed one aspect of this problem, namely by showing that provenance traces can be used to explain the differences between two sets of results that are obtained from the executions of two versions of a process [MWHW13], the latest being a reproduction of the original. Much remains to be done, however, to clearly prove the role of provenance data in data-driven, reproducible science.

References

- ADD⁺11. Yael Amsterdamer, Susan B Davidson, Daniel Deutch, Tova Milo, Julia Stoyanovich, and Val Tannen. Putting lipstick on pig: enabling database-style workflow provenance. *Proc. VLDB Endow.*, 5(4):346–357, dec 2011.
- BCBD07. O Biton, S Cohen-Boulakia, and S B Davidson. Zoom*UserViews: Querying Relevant Provenance in Workflow Systems. In *VLDB*, pages 1366–1369, 2007.
- CBL11. Sarah Cohen-Boulakia and Ulf Leser. Search, adapt, and reuse: the future of scientific workflows. *SIGMOD Rec.*, 40(2):6–16, sep 2011.

- CCT09. James Cheney, Laura Chiticariu, and Wang-Chiew Tan. Provenance in Databases: Why, How, and Where. *Foundations and Trends in Databases*, 1:379–474, 2009.
- CFLV12. James Cheney, Anthony Finkelstein, Bertram Ludaescher, and Stijn Vansummeren. Principles of Provenance (Dagstuhl Seminar 12091). *Dagstuhl Reports*, 2(2):84–113, 2012.
- CKKT11. Tyrone Cadenhead, Vaibhav Khadilkar, Murat Kantarcioglu, and Bhavani Thuraisingham. Transforming provenance using redaction. In *Proceedings of the 16th ACM symposium on Access control models and technologies, SACMAT '11*, pages 93–102, New York, NY, USA, 2011. ACM.
- CMM12. James Cheney, Paolo Missier, and Luc Moreau. Constraints of the Provenance Data Model. Technical report, 2012.
- DCBE⁺07. S Davidson, S Cohen-Boulakia, A Eyal, B Ludäscher, T McPhillips, S Bowers, M Kumar Anand, and J Freire. Provenance in Scientific Workflow Systems. In *Data Engineering Bulletin*, volume 30. dec 2007.
- DF08. S. Davidson and J. Freire. Provenance and scientific workflows: challenges and opportunities. In *Procs. SIGMOD Conference, Tutorial*, pages 1345–1350, 2008.
- DZL11. Saumen Dey, Daniel Zinn, and Bertram Ludäscher. ProPub: Towards a Declarative Approach for Publishing Customized, Policy-Aware Provenance. In Judith Bayard Cushing, James French, and Shawn Bowers, editors, *Scientific and Statistical Database Management*, volume 6809 of *Lecture Notes in Computer Science*, pages 225–243. Springer Berlin / Heidelberg, 2011.
- FM14. Hugo Firth and Paolo Missier. ProvGen: generating synthetic PROV graphs with predictable structure. In *Procs. IPAW 2014 (Provenance and Annotations)*, Koln, Germany, 2014. Springer.
- GP13. Devarshi Ghoshal and Beth Plale. Provenance from Log Files: a Big-Data Problem. In *Procs. BigProv Workshop on Managing and Querying Provenance at Scale*, 2013.
- Gre07. V. Green, Todd J., Karvounarakis, G., Tannen. Provenance Semirings. In *PODS*, pages 31–40, 2007.
- HWS⁺06. D Hull, K Wolstencroft, R Stevens, C A Goble, M R Pocock, P Li, and T Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, 34:729–732, 2006.
- HWWL11. Hugo Hiden, Paul Watson, Simon Woodman, and D. Leahy. e-Science Central: Cloud-based e-Science and its application to chemical property modelling. Technical report cs-tr-1227, School of Computing Science, Newcastle University, 2011.
- Kat14. Daniel S Katz. Transitive credit as a means to address social and technological concerns stemming from citation and attribution of digital products. *Journal of Open Research Software*, 2(1):e20, 2014.
- KS15. John E Kratz and Carly Strasser. Making data count. *Nature Scientific Data*, 2:150039, aug 2015.
- LB14a. Barbara Lerner and Emery Boose. RDataTracker: Collecting Provenance in an Interactive Scripting Environment. In *6th USENIX Workshop on the Theory and Practice of Provenance (TaPP 2014)*, 2014.
- LB14b. Barbara S Lerner and Emery R Boose. Collecting Provenance in an Interactive Scripting Environment. *Procs. TAPP'14*, 2014.

- LLCF10. Chunhyeok Lim, Shiyong Lu, A Chebotko, and F Fotouhi. Prospective and Retrospective Provenance Collection in Scientific Workflow Environments. In *Services Computing (SCC), 2010 IEEE International Conference on*, pages 449–456, jul 2010.
- LM10. John Lyle and Andrew Martin. Trusted computing and provenance: better together. In *Proceedings of the 2nd conference on Theory and practice of provenance*, TAPP’10, page 1, Berkeley, CA, USA, 2010. USENIX Association.
- LSM⁺12. Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. PROV-O: The PROV Ontology. Technical report, 2012.
- MBC⁺14. Leonardo Murta, Vanessa Braganholo, Fernando Chirigati, David Koop, and Juliana Freire. noWorkflow: Capturing and Analyzing Provenance of Scripts. *Procs. IPAW’14*, 2014.
- MBG⁺14. Paolo Missier, Jeremy Bryans, Carl Gamble, Vasa Curcin, and Roxana Danger. ProvAbs: model, policy, and tooling for abstracting PROV graphs. In *Procs. IPAW 2014 (Provenance and Annotations)*, Koln, Germany, 2014. Springer.
- MCF⁺11. Luc Moreau, Ben Clifford, Juliana Freire, Joe Futrelle, Yolanda Gil, Paul Groth, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers, Beth Plale, Yogesh Simmhan, Eric Stephan, and Jan Van Den Bussche. The Open Provenance Model — Core Specification (v1.1). *Future Generation Computer Systems*, 7(21):743–756, 2011.
- MCS11. Peter Macko, Marc Chiarini, and Margo Seltzer. Collecting Provenance via the Xen Hypervisor. In Juliana Freire and Peter Buneman, editors, *TAPP workshop*, Heraklion, Greece, 2011.
- MDB⁺13. Paolo Missier, Saumen Dey, Khalid Belhajjame, Victor Cuevas, and Bertram Ludäscher. D-PROV: extending the PROV provenance model with workflow structure. In *Procs. TAPP’13*, Lombard, IL, 2013.
- MGC⁺15. Luc Moreau, Paul Groth, James Cheney, Timothy Lebo, and Simon Miles. The rationale of PROV. *Web Semantics: Science, Services and Agents on the World Wide Web*, apr 2015.
- MHS⁺12. Luc Moreau, Olaf Hartig, Yogesh Simmhan, James Myers, Timothy Lebo, Khalid Belhajjame, and Simon Miles. PROV-AQ: Provenance Access and Query. Technical report, 2012.
- Mis15a. Editorial - Special Issue on Provenance, Data and Information Quality. *J. Data and Information Quality*, 5(3):8:1—8:4, mar 2015.
- Mis15b. Special Issue on Provenance, Data and Information Quality. *J. Data and Information Quality*, 5(3), 2015.
- MLAB08. L Moreau, B Ludäscher, I Altintas, and R S Barga. The First Provenance Challenge. *Concurrency and Computation: Practice and Experience*, 20:409–418, apr 2008.
- MMB⁺12. Luc Moreau, Paolo Missier, Khalid Belhajjame, Reza B’Far, James Cheney, Sam Coppens, Stephen Cresswell, Yolanda Gil, Paul Groth, Graham Klyne, Timothy Lebo, Jim McCusker, Simon Miles, James Myers, Satya Sahoo, and Curt Tilmes. PROV-DM: The PROV Data Model. Technical report, World Wide Web Consortium, 2012.
- MMCSR12. Luc Moreau, Paolo Missier, James Cheney, and Stian Soiland-Reyes. PROV-N: The Provenance Notation. Technical report, 2012.

- MMM05. Chris Mitchell, Chris Mitchell, and Chris Mitchell. Trusted computing. In Liqun Chen, Chris J. Mitchell, and Andrew Martin, editors, *Procs. Trust 2009*, Oxford, UK, 2005. Springer.
- MPB10. P. Missier, N. Paton, and K. Belhajjame. Fine-grained and efficient lineage querying of collection-based workflow provenance. In *Procs. EDBT*, Lausanne, Switzerland, 2010.
- MSRO⁺10. Paolo Missier, Stian Soiland-Reyes, Stuart Owen, Wei Tan, Alex Nenadic, Ian Dunlop, Alan Williams, Tom Oinn, and Carole Goble. Taverna, reloaded. In M Gertz, T Hey, and B Ludaescher, editors, *Procs. SSDBM 2010*, Heidelberg, Germany, 2010.
- MSZ⁺10. Paolo Missier, Satya S Sahoo, Jun Zhao, Amit Sheth, and Carole Goble. Janus: from Workflows to Semantic Provenance and Linked Open Data. In *Procs. IPAW 2010*, Troy, NY, 2010.
- MWHW13. Paolo Missier, Simon Woodman, Hugo Hiden, and Paul Watson. Provenance and data differencing for workflow reproducibility analysis. *Concurrency and Computation: Practice and Experience*, pages n/a—n/a, 2013.
- pro09. The Provenance Incubator Group Charter, 2009. Available at <http://www.w3.org/2005/Incubator/prov/charter>.
- pro10. The Provenance Incubator Group Final Report, 2010. Available at <http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/>.
- pro13a. PROV DC, 2013. Available at <http://www.w3.org/TR/prov-dc/>.
- pro13b. PROV Dictionary, 2013. Available at <http://www.w3.org/TR/prov-dictionary/>.
- pro13c. PROV-XML, 2013. Available at <http://www.w3.org/TR/prov-xml/>.
- pro14. The ProvONE provenance model, 2014. Available at <http://tinyurl.com/ProvONE>.
- w3c12. PROV-Overview: An Overview of the PROV Family of Documents. Technical report, 2012.
- WHW15. Simon Woodman, Hugo Hiden, and Paul Watson. Workflow Provenance: An analysis of long term storage costs. In *Procs. 10th WORKS workshop*, Austin, Texas, 2015.
- ZCL09. Jing Zhang, Adriane Chapman, and Kristen LeFevre. Do You Know Where Your Datas Been? Tamper-Evident Database Provenance. In Willem Jonker and Milan Petkovic, editors, *Secure Data Management*, volume 5776 of *Lecture Notes in Computer Science*, pages 17–32. Springer Berlin / Heidelberg, 2009.